Modelling minds as well as populations

Thomas L. Griffiths and Florencia Reali

References	**This article cites 15 articles, 6 of which can be accessed free** http://rspb.royalsocietypublishing.org/content/early/2011/03/18/rspb.2011.0349.full.html#ref-list-1
P<P	Published online 23 March 2011 in advance of the print journal.
Subject collections	Articles on similar topics can be found in the following collections cognition (492 articles)
Email alerting service	Receive free email alerts when new articles cite this article - sign up in the box at the top right-hand corner of the article or click **here**

Advance online articles have been peer reviewed and accepted for publication but have not yet appeared in the paper journal (edited, typeset versions may be posted when available prior to final publication). Advance online articles are citable and establish publication priority; they are indexed by PubMed from initial publication. Citations to Advance online articles must include the digital object identifier (DOIs) and date of initial publication.

To subscribe to *Proc. R. Soc. B* go to: **http://rspb.royalsocietypublishing.org/subscriptions**

Invited reply

Modelling minds as well as populations

Bentley *et al.* [1] present a spirited defence of the use of neutral models at the population level, and a demonstration that the three linguistic phenomena considered in our original article [2] can be captured through a different modification to the Wright–Fisher model than the one we considered. Their demonstration helps to reinforce our argument that these phenomena can be explained without the need to appeal to selection, showing that another simple neutral model can produce these effects. We see the key issues raised by their commentary as being whether there is any value in the novel connection that we identified between the Wright–Fisher model and cultural transmission by Bayesian agents, and whether this interpretation 'runs the risk of obscuring the advances made both through careful modifications and wider applications of this powerful model' (p. 1). We address these issues in turn.

To recapitulate our basic result, we showed that transmission of a probability distribution over a discrete set of alternatives by a sequence of Bayesian agents could be mathematically equivalent to the Wright–Fisher model [3,4], a classic model used in population genetics. Formally, each learner receives n observations of a set of variants (such as words or linguistic constructions), forms an estimate of the probability of each variant, summarized in a vector θ, and then generates n observations by sampling from this distribution. We showed that when the learners apply Bayesian inference with a particular prior distribution and choose the value of θ with highest posterior probability, the frequencies of the variants follow the dynamics of the Wright–Fisher model. We then used the connection to Bayesian inference to introduce a more flexible variant of this model, which we applied to the linguistic phenomena mentioned above.

We agree with Bentley *et al.* [1] that the Bayesian interpretation of this model is more complex than the original model, in which cultural transmission is described simply in terms of copying variants with some chance of 'mutation', by direct analogy to biological transmission. This complexity results from considering the cognitive mechanisms that underlie each cultural transmission event, and treating them as being more sophisticated than error-prone copying. Our interest in these mechanisms is partly a consequence of our bias as cognitive scientists, but also reflects our expectation that the mechanisms of cultural transmission are going to be different from (and potentially far more complex than) the mechanisms of biological transmission. In the case of the Wright–Fisher model, we believe that our

mathematical analysis has several implications that make this extra complexity worthwhile.

As a general methodological point, we first note that mathematical results connecting different models are valuable not just because they provide new interpretations of those models, but because they extend the tools that are available for analysing them. In this case, we provide a link to a broader class of 'iterated learning' models that have been used to model language evolution [5,6]. In these models, a sequence of agents each hear a set of utterances, form a hypothesis about the language and then generate the utterances that are heard by the next agent. When the agents use Bayesian inference, the outcome of this process is well understood. In particular, if agents choose hypotheses by sampling from their posterior distribution, over time the probability an agent selects a particular hypothesis converges to the prior probability of that hypothesis [7]. This process can be shown to be a form of Gibbs sampling, a Markov chain Monte Carlo algorithm that is widely used in Bayesian statistics [7]. These results can potentially provide new insight into the Wright–Fisher model: standard asymptotic analyses of this model use diffusion approximations (see [8]), but establishing the link to iterated learning (and Gibbs sampling) indicates that there is a closely related class of models where the asymptotic behaviour is exactly known. If agents select the hypothesis with highest posterior probability, as is required to establish equivalence to the Wright–Fisher model, then iterated learning becomes equivalent to a different statistical inference algorithm, known as the stochastic expectation-maximization algorithm [7]. Again, asymptotic analyses exist for this algorithm (e.g. [9]), and showing that Wright–Fisher is an instance of this algorithm that has the potential to allow mathematical results to generalize in both directions.

Beyond these mathematical implications, providing a link to Bayesian inference establishes a connection between existing work using the Wright–Fisher model in cultural evolution and a growing literature in cognitive science on Bayesian models of cognition. While Bentley *et al.* [1] emphasize the recent disenchantment with rational models of decision-making in behavioural economics, there has been a parallel growth of interest in rational models of cognition in cognitive psychology [10–12]. Bayesian inference provides a way to answer a key question that comes up in describing human learning and memory, indicating how the expectations of an agent combine with the observed data to yield a conclusion. The prior distribution that is used in Bayesian inference expresses those factors other than the data that influence the conclusions that agents reach, including innate dispositions, schemas established through past experience, and prior knowledge about a particular situation. In formal

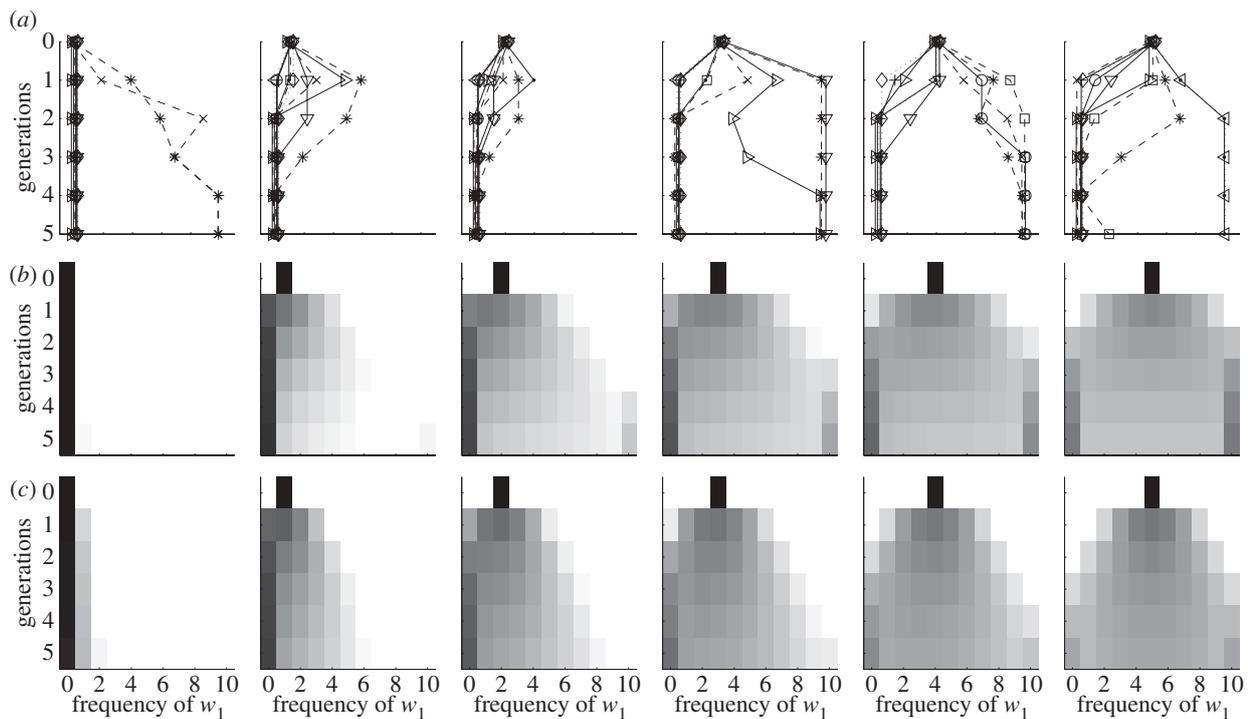The accompanying comment can be viewed at http://dx.doi.org/doi:10.1098/rspb.2010.2581.

1

Figure 1. Results of simulating cultural transmission of frequency distributions in the laboratory. In this experiment, participants were shown two labels for each of six novel objects, with the frequencies of the labels varying. Each object appeared a total of 10 times. After this initial exposure, participants produced 10 labels for each object. Participants were run in 'chains' of five, where the first participant received fixed initial frequencies, while each subsequent participant received frequencies determined by the responses of the previous participant. (*a*) Participant responses, with the frequency of one of the two labels appearing on the horizontal axis. Each chain of participants is shown with one line, with iterations shown on the vertical axis. The six conditions correspond to the different initial frequencies of the labels (in condition 0, one label never appeared and the other appeared 10 times, in condition 1, one label appeared once and the other appeared 9 times, etc.). Regardless of initial conditions, cultural transmission resulted in languages where only one label was used for each object, a form of regularization. This pattern can be modelled using the Bayesian model described in the text, with a prior that favours more deterministic languages, using either (*b*) sampling from the posterior distribution or (*c*) selection of the most probable hypothesis. Model results are shown on the same axes as in (*a*), but in greyscale where darker indicates higher probability. In this case, sampling provides a better fit to the human data. Adapted from Reali & Griffiths [14].

analyses of learning, these factors are referred to as 'inductive biases' [13]. Developing models of cultural evolution in which the mechanisms of transmission take into account the inductive biases of agents is essential if we want to understand how the structure of individual minds influences the languages and concepts adopted by societies.

Relating minds and societies also provides us with an opportunity to use cultural transmission as a tool for investigating the inductive biases of human learners. By simulating cultural transmission in the laboratory, we can discover which languages and concepts survive, and thus answer questions about human learning. In fact, we recently used the connection between Bayesian inference and the Wright–Fisher model to address a question that arises in the literature on human language acquisition [14]. Children are known to 'regularize' languages—move towards a more deterministic rule-like structure, by decreasing probabilistic variation—and this is seen as a key part of language change [15,16]. However, it is not clear whether adults share a similar tendency, with some empirical results suggesting that adults 'probability match', producing utterances with the same probabilities as seen in their linguistic input [17]. This is a question about the inductive biases of human learners—the prior distribution that people have

over probabilities in language. Our Bayesian interpretation of the Wright–Fisher model predicts how such prior distributions will affect the outcome of cultural transmission of linguistic frequency distributions. We used this connection as the basis for an experiment in which human participants simulated cultural transmission in the laboratory, learning which words to use to describe a set of six novel objects. Each object could be labelled using two different words, and the frequencies of these words varied, with each object being presented a total of 10 times. Each participant saw the objects paired with the labels with appropriate frequencies, then was shown the objects another 10 times each, being asked to produce labels for them. The results, shown in figure 1, indicated that adults do in fact have a bias towards regularization. This bias is weak enough not to be easily detected by analysing the productions of a single generation of learners, but is strong enough to influence the outcome of cultural transmission. Similar results for another linguistic construction were recently reported by Smith & Wonnacott [18].

Establishing that the Wright–Fisher model corresponds to a special case of cultural transmission by iterated learning with Bayesian agents opens the door to a deeper understanding of the differences between cultural and biological transmission. While biology
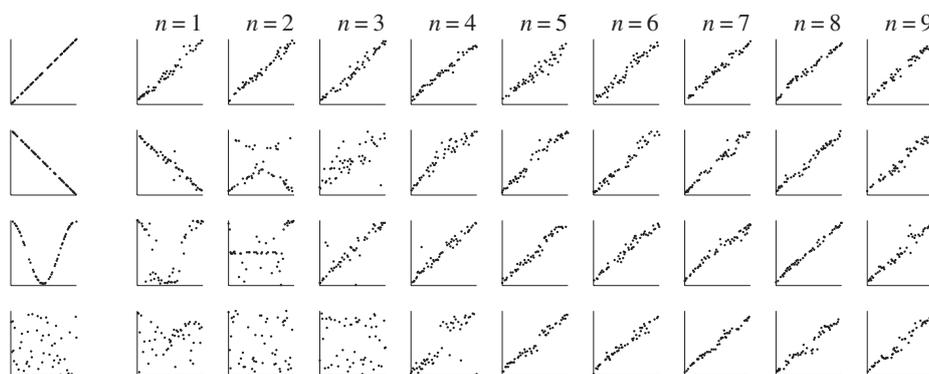
Figure 2. Cultural transmission of complex concepts reflects inductive biases. In this experiment, participants learned an abstract functional relationship between two variables, shown as the horizontal and vertical axes of these plots. Participants were trained on a set of examples of the values of the two variables, and were then asked to make predictions about the value of one variable based on the value of the other. Participants were run in 'chains' of nine, where the first participant was trained on an initial function, and each subsequent participant was trained on the predictions generated by the previous participant. The four lines of plots show four chains, with the leftmost column showing the initial training data used for the first participant, and each subsequent column showing the predictions for one of the participants in the chain (and consequently, the data used to train the next participant). Regardless of the initial data, chains converged towards a positive linear function, where both variables increase together. This is consistent with previous work suggesting that this is the easiest function for people to learn. Adapted from [19].

provides us with a well-developed account of how information changes through transmission, and it is exciting to note cases where this seems appropriate for describing phenomena we observe in human societies, copying-with-errors is clearly incomplete as an account of how people learn from one another. This problem becomes salient when we consider richer, more structured knowledge that people acquire from one another, such as languages and concepts. Modelling learning as Bayesian inference gives us the tools to predict how cultural transmission will affect these forms of knowledge, and how inductive biases are involved. Laboratory simulations of cultural transmission of different forms of knowledge confirm the prediction that such knowledge will change to take forms that are consistent with human inductive biases [19]. An example is shown in figure 2, where a functional relationship between two variables changes to take the form that people find easiest to learn through a few generations of cultural transmission [19]. By placing the Wright–Fisher model in this broader framework, we can show that the mechanisms of cultural transmission are strictly more general than those of biological transmission.

Turning to the question of whether other modelling advances might be obscured, we see no reason why the Bayesian interpretation of the Wright–Fisher model we proposed presents a challenge to previous 'population-level' applications of this model. Providing a new interpretation for a model does not invalidate results that interpret it differently, and the structure of the Wright–Fisher model is sufficiently simple that we might expect to see similar dynamics resulting from many different mechanisms of transmission (including biological reproduction, copying-with-errors and Bayesian inference). However, it is important to recognize that these different interpretations turn primarily on the mechanisms of transmission, and not whether the model is applied to populations or individuals, as Bentley *et al.* [1,5] suggest. It is possible to define population-level models in which individual learners use Bayesian

inference and obtain exactly the same mathematical results as for our individual-level model. For example, we might have a population of n agents, each of whom observes the variants generated by the n agents in the previous generation, applies Bayesian inference as specified in our model, and then generates a single variant by sampling from the resulting distribution.

Ultimately, we expect that progress in modelling cultural evolution will result from recognizing that different transmission mechanisms are going to be relevant in different contexts. The copying-with-errors (or innovations) mechanism emphasized by Bentley *et al.* [1] is reasonable in many of the settings in which it has been applied in their previous work, such as aesthetic choices about baby names or ceramic styles [20,21]. Bayesian inference is appropriate for describing more complex instances of learning or memory, as supported by the use of these models to describe such processes in cognitive science. These mechanisms happen to converge in the Wright–Fisher model, but this simple model is just a starting point for exploring the processes that are involved in cultural evolution. We hope that this starting point is used to begin a deeper exploration of how models of individual minds can be connected to models that describe the behaviour of populations.

Thomas L. Griffiths and Florencia Reali*
Department of Psychology, University of California, 3210 Tolman Hall, MC 1650, Berkeley, CA 94720-1650, USA
*Author for correspondence (florencia.reali@gmail.com).

REFERENCES

1 Bentley, R., Ormerod, P. & Shennan, S. J. In press. Population-level neutral model already explains linguistic patterns. *Proc. R. Soc. B.* (doi:10.1098/rspb.2010.2581)

2 Reali, F. & Griffiths, T. L. 2010 Words as alleles: connecting language evolution with Bayesian learners to models of genetic drift. *Proc. R. Soc. B* **277**, 429–436. (doi:10.1098/rspb.2009.1513)

3 Fisher, R. A. 1930 *The genetical theory of natural selection*. Oxford, UK: Clarendon Press.

4 Wright, S. 1931 Evolution in Mendelian populations. *Genetics* **16**, 97–159.

5 Kirby, S. 2001 Spontaneous evolution of linguistic structure: an iterated learning model of the emergence of regularity and irregularity. *IEEE J. Evol. Comput.* **5**, 102–110. (doi:10.1109/4235.918430)

6 Kirby, S., Dowman, M. & Griffiths, T. L. 2007 Innateness and culture in the evolution of language. *Proc. Natl Acad. Sci. USA* **104**, 5241–5245. (doi:10.1073/pnas.0608222104)

7 Griffiths, T. L. & Kalish, M. L. 2007 Language evolution by iterated learning with Bayesian agents. *Cogn. Sci.* **31**, 441–480. (doi:10.1080/15326900701326576)

8 Ewens, W. J. 2004 *Mathematical population genetics*. New York, NY: Springer-Verlag.

9 Nielsen, S. F. 2000 The stochastic EM algorithm: estimation and asymptotic results. *Bernoulli* **6**, 457–489. (doi:10.2307/3318671)

10 Anderson, J. R. 1990 *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

11 Chater, N. & Oaksford, M. 2008 *The probabilistic mind: prospects for Bayesian cognitive science*. Oxford, UK: Oxford University Press.

12 Tenenbaum, J. B., Griffiths, T. L. & Kemp, C. 2006 Theory-based Bayesian models of inductive learning and reasoning. *Trends Cogn. Sci.* **10**, 309–318. (doi:10.1016/j.tics.2006.05.009)

13 Mitchell, T. 1997 *Machine learning*. New York, NY: McGraw Hill.

14 Reali, F. & Griffiths, T. L. 2009 The evolution of frequency distributions: relating regularization to inductive biases through iterated learning. *Cognition* **111**, 317–328. (doi:10.1016/j.cognition.2009.02.012)

15 Bickerton, D. 1981 *Roots of language*. Ann Arbor, MI: Karoma.

16 Hudson, K. C. & Newport, E. 2005 Regularizing unpredictable variation: the roles of adult and child learners in language formation and change. *Lang. Learn. Dev.* **1**, 151–195. (doi:10.1207/s15473341lld0102_3)

17 Vouloumanos, A. 2008 Fine-grained sensitivity to statistical information in adult word learning. *Cognition* **107**, 729–742. (doi:10.1016/j.cognition.2007.08.007)

18 Smith, K. & Wonnacott, E. 2010 Eliminating unpredictable variation through iterated learning. *Cognition* **116**, 444–449. (doi:10.1016/j.cognition.2010.06.004)

19 Griffiths, T. L., Kalish, M. L. & Lewandowsky, S. 2008 Theoretical and experimental evidence for the impact of inductive biases on cultural evolution. *Phil. Trans. R. Soc. B* **363**, 3503–3514. (doi:10.1098/rstb.2008.0146)

20 Hahn, M. W. & Bentley, R. 2003 Drift as a mechanism for cultural change: an example from baby names. *Proc. R. Soc. Lond. B* **270**, S120–S123. (doi:10.1098/rsbl.2003.0045)

21 Bentley, R., Hahn, M. W. & Shennan, S. J. 2004 Random drift and culture change. *Proc. R. Soc. Lond. B* **271**, 1143–1150. (doi:10.1098/rspb.2004.2746)